

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 4 年 2 月 2 4 日
Date of Application:

出 願 番 号 特 願 2 0 0 4 - 0 4 7 1 7 6
Application Number:
[ST. 10/C]: [J P 2 0 0 4 - 0 4 7 1 7 6]

出 願 人 株式会社日立製作所
Applicant(s):

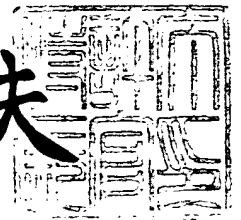
CERTIFIED COPY OF
PRIORITY DOCUMENT

BEST AVAILABLE COPY

2 0 0 4 年 3 月 2 4 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 4 - 3 0 2 4 1 8 6

【書類名】 特許願
【整理番号】 K03011801A
【あて先】 特許庁長官殿
【国際特許分類】 G06F 12/00
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 出射 英臣
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 西川 記史
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 茂木 和彦
【特許出願人】
 【識別番号】 000005108
 【氏名又は名称】 株式会社 日立製作所
【代理人】
 【識別番号】 100075096
 【弁理士】
 【氏名又は名称】 作田 康夫
【選任した代理人】
 【識別番号】 100100310
 【弁理士】
 【氏名又は名称】 井上 学
【国等の委託研究の成果に係る記載事項】 国等の委託研究の成果に係る特許出願（平成
1 5 年度 文部科学省 科学技術試験研究委託費〔1〕差分デー
タ及びログ転送によるリモートサーバレス化技術の実現方式検討
、2）クエリプラン利用先読み技術の実現方式検討、3）モニタ
情報に基づくデータベース管理の自律化技術の実現方式検討）（
産業活力再生特別措置法第 3 0 条の適用を受けるもの）
【手数料の表示】
 【予納台帳番号】 013088
 【納付金額】 21,000円
【提出物件の目録】
 【物件名】 特許請求の範囲 1
 【物件名】 明細書 1
 【物件名】 図面 1
 【物件名】 要約書 1

【書類名】 特許請求の範囲**【請求項 1】**

第一の計算機及び第一のストレージ装置を有する第一のサイト、
第二の計算機及び第二のストレージ装置を有する第二のサイト、及び
管理用の計算機、前記第一のサイト、前記第二のサイト及び前記管理用の計算機とを相互に接続するネットワークとを有し、

前記第一のストレージ装置は、前記管理用の計算機に入力された情報に基づいて該ストレージ装置が有する記憶領域に格納されるデータをグルーピングし、前記グルーピングされたグループ単位で該グループ内で更新されたデータを前記第二のストレージ装置へ転送し、

前記第一のサイトが停止した際に、

前記第二のサイトが前記グループ単位にデータを復旧することを特徴とする計算機システム。

【請求項 2】

前記第一のストレージ装置は、

前記第二のサイトでデータを復旧する際に要求される復旧時間に基づいて優先順位をつけて前記グルーピングを行い、

前記第二のサイトは、

前記要求される復旧時間の優先順位が高い順に前記グルーピングされたグループに含まれるデータを復旧することを特徴とする請求項 1 記載の計算機システム。

【請求項 3】

前記第一のストレージ装置から前記第二のストレージ装置への前記グループ単位のデータ転送は非同期リモートコピーで行われることを特徴とする請求項 2 記載の計算機システム。

【請求項 4】

前記グルーピングされたデータは、データベースに使用されるデータであることを特徴とする請求項 3 記載の計算機システム。

【請求項 5】

前記グルーピングされるデータには、前記データベースで使用されるログデータが含まれており、

前記ログデータが含まれるグループのデータは、同期リモートコピーで前記第一のストレージ装置から前記第二のストレージ装置へ転送されることを特徴とする請求項 4 記載の計算機システム。

【請求項 6】

前記ログデータが含まれるグループには一番高い復旧時間の優先順位が割り当てられていることを特徴とする請求項 5 記載の計算機システム。

【請求項 7】

前記第二のサイトは、前記第一のサイトの停止を前記ネットワークを介して検知することを特徴とする請求項 6 記載の計算機システム。

【請求項 8】

前記第二のサイトは、

前記グループ単位でデータを復旧する際に、復旧前のデータが含まれているグループの使用を禁止し、グループ単位でのデータの復旧が完了するたびに復旧されたデータが含まれるグループの使用を許可することを特徴とする請求項 7 記載の計算機システム。

【請求項 9】

計算機及び管理用計算機と接続されるストレージ装置であって、

他のストレージ装置とネットワークを介して接続され、

制御部及びメモリを有し、

前記制御部は、前記管理用計算機に入力された情報に基づいて該ストレージ装置が有する記憶領域に格納されるデータをグルーピングし、前記グルーピングの情報を前記メモリ

に格納し、

前記メモリに格納された情報に基づいて、前記グルーピングされたグループ単位で該グループ内で更新されたデータを前記第二のストレージ装置へ転送することを特徴とするストレージ装置。

【請求項 10】

前記グルーピングは、前記他のストレージ装置におけるデータの復旧時間に基づいた優先順位に基づいて行われることを特徴とする請求項 9 記載のストレージ装置。

【請求項 11】

第一の計算機及び第一のストレージ装置を有する第一のサイト、

第二の計算機及び第二のストレージ装置を有する第二のサイト、及び

管理用の計算機、前記第一のサイト、前記第二のサイト及び前記管理用の計算機とを相互に接続するネットワークとを有し、

前記第一のストレージ装置は、前記管理用の計算機に入力された前記第二のサイトでデータを復旧する際に要求される復旧時間に基づいた優先順位に基づいてストレージ装置が有する記憶領域に格納されるデータをグルーピングし、前記グルーピングされたグループ単位で該グループ内で更新されたデータを前記第二のストレージ装置へ転送し、

前記第一のサイトが停止した際に、

前記第二のサイトは、

前記要求される復旧時間の優先順位が高い順に前記グルーピングされたグループに含まれるデータを復旧することを特徴とする計算機システム。

【書類名】明細書

【発明の名称】計算機システム

【技術分野】

【0001】

本発明は、計算機システム、特にデータベースを運用する計算機システム（以下「DBシステム」）において、通常運用されているDBシステム（以下「現用系システム」）と、現用系システムのデータが複製されて使用される計算機システム（以下「待機系システム」）間におけるデータコピー方法及び待機系システムにおけるデータ回復の方法に関する。

【背景技術】

【0002】

データベース（以下「DB」とも称する）が構築され、そのDBを運用しているDBシステムにおいて、DBを構成するデータ（以下「DBデータ」）のバックアップを取る方法として、現用系システムと同じ構成のDBシステムを待機系システムとして用意し、現用系システムのDBデータを待機系システムにコピーする技術がある。

【0003】

上記方法を用いれば、現用系システムが何らかの理由により稼働を停止した際、待機系システムを稼働させることでシステム全体の可用性が向上する。また、現用系システムと待機系システムとを物理的に距離が離れた場所に各々設置することで、地震等の自然災害が現用系システムを設置した場所に発生した場合でも、待機系システムでDBデータを回復することが可能となる。以下、このような現用系、待機系システムを有するシステムをディザスタリカバリ（DR）システムと呼ぶ。

【0004】

DRシステムにおいて、現用系システムと待機系システムとの間のデータ転送が、DBデータを格納するストレージ装置間で行われる場合がある。このようなストレージ装置間のデータ転送を以下リモートコピーと称する。

【0005】

リモートコピーの手法として、同期リモートコピーと非同期リモートコピーの2種類がある。同期リモートコピーでは、現用系システムのストレージ装置と待機系システムのストレージ装置間でDBデータの更新の同期を取る（DBデータの内容の同一性を保証するため、待機系システムに転送されたデータの信頼性は高い。一方、DBデータの同期をとるため、現用系システムはDBデータを待機系システムへ転送している間処理の終了の報告を待つ必要があり、現用系システムの性能が低下する。

【0006】

一方、非同期リモートコピーでは、現用系システムのストレージ装置と待機系システムのストレージ装置間でDBデータの同期を取らずにデータを転送するため、同期リモートコピーとは逆に、現用系システムの性能はそれほど低下しないが、待機系システムに転送されたデータの信頼性（あるいは最新性）は低くなる。

【0007】

特許文献1では、現用系と待機系のストレージ装置間において、論理ボリュームグループ毎に優先度をつけ、その優先度に従って非同期リモートコピーを実行する方法を開示している。また、その一例としてDBのログが記憶される論理ボリュームの優先度を高くし、他のDBデータと比して、ログを優先的に待機系システムのストレージ装置へ転送することが開示されている。

【0008】

【特許文献1】特開2003-6016号公報

【発明の開示】

【発明が解決しようとする課題】

【0009】

上述のDRシステムを構築する際、DBデータの種類毎に、待機系システムにおけるD

Bデータの復旧に掛けられる時間が異なる場合がある。しかし従来技術ではその復旧時間の差が考慮されておらず、通常、最優先に復旧処理を行わなければならないDBデータの復旧時間内で全てのDBデータの復旧処理を行うようなDRシステムが構築される。しかしこれではDRシステムがオーバースペックになる可能性が高く、最適なシステム構成とはならない。

【0010】

本発明の目的は、最優先に復旧すべきデータ、例えばDBデータの待機系システムにおける復旧時間を短く保ちながら、効果対価格比の良いシステムを構築することにある。

【課題を解決するための手段】

【0011】

上記目的を達成するため、本発明では以下の構成を採用する。具体的には、正サイトから副サイトへデータをリモートコピーする際に、予め副サイトにおいて要求されるデータの復旧時間に応じて正サイトでデータを区別（グループ化）し、その区別に応じてデータを転送する順序を決定し、正サイトがリモートコピーを行う。その後、副サイトにおいては、正サイトの障害発生時に、上述したグループに基づいてデータの復旧を行う構成とする。

【0012】

データのグループ化を行う際には、副サイトにおいてデータ復旧に要求される時間が短い順にグループ化を行い、正サイトはそのグループに基づいてデータを転送し、かつ副サイトにおいてデータ復旧にかけられる時間が少ないグループ順に復旧を行う。

【0013】

さらに、データ復旧にかけられる時間の多少に関わらず、データベースにおけるログデータについては最先にデータを副サイトに転送してもよい。さらにこの場合、同期リモートコピーを用いてログデータを副サイトに転送しても良い。

【0014】

又、副サイトにおいてデータを復旧する際に、一旦全てのデータを閉塞し、復旧されたグループ単位で閉塞を解除しても良い。

【発明の効果】

【0015】

本発明によれば、待機系システムにおいて重要度の高いデータの復旧時間を短くしながらより最適なDRシステムを構築することが可能となる。

【発明を実施するための最良の形態】

【0016】

以下、図1～図14を用いて本発明の実施の形態を説明する。なお、これにより本発明が限定されるものではない。

図1は、本発明を適用した計算機システムの実施形態の例を示す図である。

計算機システムは、現用系システムである正サイト100、待機系システムである副サイト102、正サイト100と副サイト102とを接続するネットワーク172及び正サイト100と副サイト102を管理するために使用されるシステム管理サーバ150を有する。

【0017】

正サイト100及び副サイト102は、データベースマネジメントシステム（以下、DBMS）が動作するDBサーバ110、DBデータを記憶するストレージ装置130並びにDBサーバ110及びストレージ装置130を相互に接続するネットワーク170を有する。以下、正サイト100と副サイト102の機器構成は同等として説明するが、全く同じでなくても良い。少なくとも双方のサイトでDBサーバとストレージ装置があれば良い。

【0018】

DBサーバ110は一般的な計算機であり、ユーザが使用する計算機もしくは当該計算機が接続されるネットワークとのインタフェースであるI/F(A)112、制御装置（

制御プロセッサ) 114、入出力装置 116、メモリ 118、ネットワーク 170 とのインタフェースである I/F (B) 124 を有する。また、メモリ 118 には、運用管理プログラム 120 及び DBMS を実行するプログラム (以下「DBMS」と略する) 122 が格納される。

【0019】

ストレージ装置 130 は、ネットワーク 170 とのインタフェースである I/F 132、制御装置 (制御プロセッサ) 134、メモリ 136 及びディスク 144 を有する。また、メモリ 136 には、制御プログラム 138、リモートコピー (以下「RC」) を実行するプログラム 140 及び制御情報 142 が格納される。尚、ディスク 144 は記憶媒体であり、ハードディスクドライブでも、光ディスクでも良い。また、ストレージ装置 130 は、複数のディスク 144 を RAID 構成にしている。

【0020】

計算機システムの管理者は、正サイト 100 内のネットワーク 170 に接続されたシステム管理サーバ 150 により、計算機システムの設定・管理を行う。

【0021】

システム管理サーバ 150 は一般的な計算機であり、制御装置 (制御プロセッサ) 152、入出力装置 154、メモリ 158 及び正サイトのネットワーク 170 とのインタフェースである I/F 160 を有する。なお、システム管理サーバ 150 と正サイトの DB サーバ 110 とを一体とした構成も考えられる。この場合、管理プログラム 158 が DB サーバ 110 で実行される。

【0022】

又、上述した各種プログラムは、光ディスクやフロッピー (登録商標) ディスク等の記憶媒体を用いて各サイトにインストールされたり、ネットワークを介して各サイトの各装置にインストールされる。尚、以下でプログラムが主語になる場合は、実際はそのプログラムを実行する制御装置が処理を実行している。

【0023】

図 2 の (a) は、ストレージ装置 130 が持つ制御情報 142 の構成例を示した図である。

制御情報 142 は、ストレージ装置 130 が有する論理的な記憶領域 (論理ユニット: LU) の設定に関する情報である LU 情報 200 及びデータの優先度毎にグルーピングされるグループの情報であるグループ情報 202 を有する。尚、LU は、一つ又はそれ以上のディスク 144 から構成される。

【0024】

図 2 の (b) は、LU 情報 200 の構成例を示す図である。LU 情報 200 は、各 LU 毎に対応するエントリを有する。各エントリは、エントリに対応する LU を識別する番号である LUN を登録するフィールド 210、対応する LU に割当てられたブロックの数を示すブロック数を登録するフィールド 212 及び対応する LU が割当てられる DB サーバ 110 を識別するためのサーバ ID を登録するフィールド 214 を有する。なお、ブロックとは記憶領域を扱う際の単位を示し、通常 512KB である。

【0025】

図 2 の (c) は、グループ情報 202 の構成例を示す図である。グループ情報 202 は、優先度のグループ毎に対応するエントリを有する。各エントリは、エントリに対応するグループを識別するグループ ID が登録されるフィールド 220、対応するグループ内のデータの優先度を示すデータ優先度が登録されるフィールド 222 及び対応するグループに属する LU を示す LUN を登録するフィールド 224 を有する。

【0026】

図 3 は、運用管理プログラム 120 が保持するローデバース情報 300 の例を示した図である。このローデバース情報 300 は、DB サーバ 110 のメモリ 118 に格納されている。

【0027】

ローデバイス情報 300 は、DBサーバ 100 に登録されるローデバイス毎に対応するエントリを有する。ここでローデバイスとは、DBサーバ 100 で実行されるオペレーティングシステム、特にファイルシステムにおいて認識される仮想的な外部装置を指す。ファイルシステムはこのローデバイスを 1 つのファイルとして表現して操作する。これをローデバイスファイルと言う。

各エントリは、該エントリに対応するローデバイスファイルのファイル名を登録するフィールド 302、ローデバイスファイルで表現されるローデバイスと関連付けられる LU を有するストレージ装置 130 の識別子を登録するフィールド 304 及びローデバイスに対応する LU の LUN を登録するフィールド 306 とを有する。

【0028】

図 4 の (a) は、DBMS 122 の構成例を示した図である。

DBMS 122 は、DBMS 実行プログラム 400、DB データの読み出しや書き込みの際に DB データを一時保持する DB バッファ 402、DB データの更新に伴って追加されるログを一時保持するログバッファ 404、DBMS 122 のシステム情報やログ、DB データ等を記憶させるデータ領域の設定情報であるデータ領域設定情報 406 及び表や索引といった DBMS 122 のスキーマ情報である DB スキーマ情報 408 を有する。

【0029】

図 4 の (b) は、データ領域設定情報 406 の構成例を示す図である。データ領域設定情報 406 は、データ領域毎に対応するエントリを有する。各エントリは、該エントリに対応するデータ領域を識別するためのデータ領域 ID を登録するフィールド 410、対応するデータ領域の名前を示すデータ領域名を登録するフィールド 412、対応するデータ領域が作成されるローデバイスファイルのファイル名が登録されるフィールド 414、対応するデータ領域に格納されるデータの種別を示すデータ種別が登録されるフィールド 416、対応するデータ領域の大きさを示す領域サイズが登録されるフィールド 418 及び対応するデータ領域に格納されるデータの優先度を示すデータ優先度を登録するフィールド 420 を有する。

【0030】

図 4 の (c) は、DB スキーマ情報 408 の構成例を示す図である。DB スキーマ情報 408 は、DB スキーマ毎に対応するエントリを有する。各エントリは、エントリに対応する DB スキーマを識別するためのスキーマ ID を登録するフィールド 430、対応する DB スキーマの名前を示すスキーマ名を登録するフィールド 432、対応する DB スキーマの種別を示すスキーマ種別を登録するフィールド 434、対応する DB スキーマが作られるデータ領域を識別するデータ領域 ID を登録するフィールド 436 及び対応する DB スキーマに割り当てている記憶領域のデータサイズを示すスキーマサイズを登録するフィールド 438 を有する。

【0031】

なお、スキーマ種別が“TABLE”の場合は DB の表データ、“INDEX”の場合は DB の索引データがそのスキーマ内に格納される。

【0032】

図 5 は、DBMS 122 によってストレージ装置 130 に作成されるデータ領域の構成例を示した図である。上述したように、データ領域はデバイスファイル名を介してストレージ装置 130 が有する各 LU と対応付けられる。

データ領域 500 は、データ領域の管理情報を格納するデータ格納領域管理領域 502 及び実際の DB スキーマのデータを格納するデータ格納領域 504 を有する。データ領域管理領域 502 は、データ領域を識別するためのデータ領域 ID 510 を格納する領域を有する。

【0033】

図 5 (a) は、データ領域 500 にログデータを格納した例を示した図である。ログデータはデータ格納領域 504 に格納される。ログデータが格納される領域 520 には、複数のログエントリが格納される。一つのログエントリの領域は、対応するログエントリを

識別するために付与されるログシーケンス番号（以下、LSN）を格納する領域530、対応するログの種類を識別するログ種別を格納する領域532、追加更新されたデータが記憶されているデータ領域を示すデータ領域IDが格納される領域534、ログエントリに対応するデータのアドレスが格納される領域536及び追加更新されたデータ内容が格納される領域538を有する。

【0034】

このログデータの1つ1つのエントリは、DBへデータが追加更新された時や、DBがコミットされた際にDBMS122によって自動的に作成される。作成された後、このログエントリは一旦ログバッファ404に保持される。そして、ログバッファ404の記憶容量が不足した時や、ある決められた時間が経過した時、DBMS122はログバッファ404に保持されているログデータをストレージ装置130のデータ領域500に書き込む。

【0035】

図5（b）は、データ領域にDBデータを格納した場合を示した図である。DBデータはデータ格納領域504に格納される。DBMS122がDBデータを追加又は更新した際、追加更新のDBデータは一旦DBバッファ402に保持される。そして、DBバッファの記憶容量が不足した時や、ある決められた時間が経過した時、DBMS122はDBバッファ402に保持されているDBデータをストレージ装置130のデータ領域に書き込む。

【0036】

また、データ領域管理領域502は、そのデータ領域内で最も最近更新されたデータのLSNを格納する領域512を有する。例えば、あるデータ領域のデータが更新された時に作られたログデータのLSNが10の場合、この10という値が領域512に格納される。これにより、DBサーバ110が、LSN10までのログデータが当データ領域に反映されていることを判断することが可能となる。

【0037】

図6、図7は、計算機システムにおける正サイト100から副サイト102へのデータ転送の概略を示した図である。

ストレージ装置130は、RCプログラム140を実行して、正サイト100から副サイト102へデータをコピーする。このリモートコピーは、グループ単位で優先度順に行われる。また、優先度（「データ優先度」とも言う）が0であるグループのデータ（本実施形態ではログデータ）は同期リモートコピー、データ優先度が0以外のグループのデータ（本実施の形態ではDBデータ）は非同期リモートコピーで転送される。

【0038】

図6において、正サイト100のグループ0（600）はデータ優先度が0であるため、このグループ0（600）に属するLU内のデータが更新される度に、正サイト100のストレージ装置130は、同期リモートコピー（矢印610）によって副サイト102のストレージ装置130にデータを転送する。

【0039】

一方、図6においてグループ1（602）、グループ2（604）、グループ3（606）のデータ優先度は、それぞれ1、5、3である。従って、これらのグループに属するLU内のデータを副サイト102にコピーする際、正サイト100のストレージ装置130は、グループ1（602）に属するLUのデータ、グループ3（606）に属するLUのデータ、グループ2（604）に属するLUのデータの順番で非同期リモートコピー（矢印612、614、616）で副サイト102のストレージ装置130へデータ転送を行う。

【0040】

図7は、データ領域の観点から見た正副サイト間のデータ転送の概略を示した図である。以下、正サイト及び副サイトのストレージ装置130に、グループ1、2及び3に対応するデータ領域700、702及び704が構成されているとする。尚、データ領域70

0等は上述したデータ領域500の具体例であるとする。又、グループは同じ優先度に属するLUで構成される。

DBMS122によって、データ領域700にデータ「A」、データ領域702にデータ「B」、データ領域704にデータ「C」、データ領域700にデータ「D」、データ領域702にデータ「E」及びデータ領域704にデータ「F」という順番でデータ更新が行われた場合、図7のログデータ710（データ領域500の一種）に示されるログが作られる。

【0041】

ログデータ710を格納するデータ領域は、データ優先度が0のグループに属しているLUに作られているため、データの書き込みがある度にそれらのデータは同期リモートコピー（矢印610）で副サイト102へ転送される。

【0042】

データ領域700、データ領域702及びデータ領域704は、それぞれ前述のグループ1（602）、グループ2（604）及びグループ3（606）に属しているLUに作られているため、データ優先度に従い、データ領域700のデータ「A」「D」、データ領域704のデータ「B」「E」、データ領域702のデータ「C」「F」という順番で、非同期リモートコピーで正サイト100から副サイト102にデータが転送される。また、データが正サイト100から副サイト102へコピーされた際は、そのデータに対応するログのLSNも、副サイト102のデータ領域管理領域502のLSN512に転送される。

【0043】

図8～図14は、計算機システムにおける、記憶領域の設定処理から災害発生によるシステム回復処理までの一連の処理手順の例を示したフロー図である。

簡略に処理手順を説明すると、まず、計算機システムの管理者等は計算機システムが使用する記憶領域の設定を行う。この際、管理者等は副サイトでのデータの復旧時間に基づく優先順位に基づいて、記憶領域の優先度を決め、その情報を正ストレージ装置や副サイトへ登録する。その後、正副サイトは、設定された優先度に基づいてデータのリモートコピーを行う。

【0044】

正サイトで障害が発生した場合、管理者は副サイトでシステムの復旧を図る。この際、副サイトの計算機システムは、あらかじめ設定された優先度の順番にデータの復旧処理を行う。この際、副サイトのストレージ装置は、一旦すべての記憶領域の使用を禁止し、その後、復旧が完了した記憶領域から順に使用禁止を解除していく。以下、各段階での処理手順を詳細に説明する。

【0045】

図8は、記憶領域の割当て処理の手順例を示したフロー図である。

計算機システムの管理者は、計算機システムを構築するにあたり、まずシステム管理サーバ150の管理プログラム158を起動し、正サイト100のストレージ装置130内のLUを設定する。ここで設定された内容は、ネットワークを介してストレージ装置130が有するLU情報200に格納される。具体的には、例えばLU情報200のエントリ216は、管理者がLUNが0のLUに20971520ブロックを割り当て、かつサーバIDが0のDBサーバ110にLUN0のLUを割り当てたことを示している（ステップ802）。

【0046】

続いて、システム管理者は、正サイト100のDBサーバ110上のローデバイスを設定する。ここで設定された内容は、ネットワークを介してDBサーバ110が有するローデバイス情報300に格納される。具体的には、例えばローデバイス情報300のエントリ308は、管理者が“/dev/rdisk/c0t0d1s6”というファイル名のローデバイスに、ストレージ装置ID0のストレージ装置130が有するLUN0のLUを対応付けた（以下「マッピング」）ことを示している。なお、ローデバイスにマッピング可能なLUは、ステ

ップ802で設定したLUである（ステップ804）。

【0047】

続いて管理者は、副サイト102のストレージ装置130内のLUを設定する。尚、設定の内容は、サーバID以外ステップ802で設定した内容と同じにする必要がある。これは、リモートコピー時の正副サイト間の整合性をとるためである（ステップ806）。

【0048】

続いて管理者は、副サイト102のDBサーバ110のローデバイスを設定する。設定の内容は、ストレージ装置ID以外ステップ804で設定した内容と同じにする必要がある。これも、正副サイト間の整合性を取るためである（ステップ806）。

【0049】

図9は、計算機システムにおけるDBシステムの構築処理の手順例を示したフロー図である。本処理は、記憶領域の割り当て処理の後に行われる。

DBの構築を行うDB設計者等は、正サイト100のDBMS122を起動し、データ領域の名前、データ領域を作成するローデバイスのローデバイスファイル名、データ領域に格納するデータの種別、データ領域のサイズ、格納するデータの優先度を指定したクエリを入出力装置116から入力してデータ領域を作成する。

【0050】

データの優先度は具体的には以下のように決定する。まず、ログデータは副サイト102におけるデータ復旧に欠かせないため、ログデータを格納するデータ領域のデータ優先度は最も高く設定する。それ以外のDBデータは、データの重要度やDB処理の内容によって、障害発生時、副サイト102において早く復旧してほしい順にデータ優先度を決定する。例えば、常々更新処理が行われるオンライン系のDBデータは一刻も早く復旧して業務を再開する必要があるため、データ優先度を高く設定する。夜間などに一括して処理するバッチ系のDBデータは決められた時間内で処理されるため、優先度は中ほどに設定する。アーカイブ系のDBは検索／更新処理が頻繁に発生しないため、優先度は低く設定する。

【0051】

データ領域を作成した時点で、DBMS122は、そのデータ領域に識別子（ID）を自動的に割り振る。ここで設定された内容は、DBMS122内のデータ領域設定情報406に格納される。具体的には、例えばデータ領域設定情報406のエントリ422は、データ領域IDに0が割り振られたデータ領域の領域名は“LOG”であり、“/dev/rds/k/c0t0d1s6”というローデバイスファイル上に作られていることを示す。又、このデータ領域に格納されるデータはDBMS122のログデータで、領域の大きさは4GB、データの優先度は0であることを示している（ステップ902）。

【0052】

ステップ902でデータ領域が作られた後、DBサーバ110はデータ領域設定完了通知と一緒にローデバイス情報300及びデータ領域設定情報406をストレージ装置130に送信する。ストレージ装置130は、DBサーバ110からデータ領域設定完了通知、ローデバイス情報300及びデータ領域設定情報406を受信した後に、制御プログラム138のグループ作成処理1000を実行する。グループ作成処理1000については後述する。（ステップ904）。

【0053】

続いて、DB作成者は、ステップ902で作成したデータ領域上に、DBスキーマの名前、種別、DBスキーマを作成するデータ領域のID、DBスキーマのサイズをDBMS122に指定してDBスキーマの作成を指示する。DBスキーマを作成した時点で、DBMS122はそのDBスキーマにスキーマIDを自動的に割り振る。ここで設定された内容は、DBMS122内のDBスキーマ情報408に格納される。

【0054】

具体的には、例えばDBスキーマ情報408のエントリ440は、DBスキーマID0のスキーマ名は“LOG1”であり、データ領域IDが0のデータ領域に作られているこ

とを示している。又、このDBスキーマはDBMSのログデータを格納するスキーマであり、スキーマの大きさは2GBであることを示している。

尚、一つのDBMS122によってログが複数発生する場合、例えばDBMS122が2つ以上のログを切り替えて使用するといった場合は、ログ用のデータ領域を複数作成することで対応する(ステップ906)。

【0055】

続いて、DB作成者は、副サイト102のDBMS122を起動し、ステップ902で作成したデータ領域と同じデータ領域を副サイト102にも作成する(ステップ908)。

【0056】

ステップ908で副サイト102にデータ領域が作られた後、副サイト102のストレージ装置130は、制御プログラム138のグループ作成処理1000を実行する。グループ作成処理1000については後述する(ステップ910)。

【0057】

続いて、DB作成者は、ステップ908で作成したデータ領域に、ステップ906で作成したDBスキーマと同じDBスキーマを作成する。なお、ステップ908から本ステップにおいて正サイト100と同じ処理を繰り返す代わりに、正サイト100のデータ領域設定情報406や、DBスキーマ情報408を副サイト102へコピーして、正サイト100と同じDBを副サイト102上に構築しても良い(ステップ912)。

その後、正サイト100のDBMS122の運用を開始する(ステップ914)。

【0058】

図10は、制御プログラム138によるグループ作成処理1000の処理手順の例を示したフロー図である。なお、上述したように本処理はデータ領域が作成されたときに実行される。

ストレージ装置130は、制御プログラム138を実行して、作られたデータ領域のデータ優先度とそのデータ領域が属するLUのLUNについての情報をDBサーバ110のローデバース情報300及びデータ領域設定情報406から取得し、そのLUNに対応するLUのデータ優先度を決定する(ステップ1002)。

【0059】

続いて、ストレージ装置130は、ステップ1002で取得したデータ優先度と同じデータ優先度のグループがあるか、グループ情報202を検索する(ステップ1004)。

【0060】

ステップ1004で同じデータ優先度のグループがあった場合、ストレージ装置130は、同じデータ優先度のグループにステップ1002で取得したLUNを追加し、グループ情報202を更新する(ステップ1006)。

【0061】

一方、ステップ1005で同じデータ優先度のグループがなかった場合、ストレージ装置130は、グループ情報202に新たなエントリを追加することで新しいグループを作成し、ステップ1002で取得したLUNをそのエントリのフィールド244に追加する(ステップ1008)。

【0062】

上述のステップによって、グループ情報202が更新される。具体的には、例えばグループ情報202のエントリ226は、グループIDが0のグループ内のデータ優先度は0であり、LUN0のLUが当グループに属していることを示している。

【0063】

図11は、DBMS122によるデータ更新処理の手順例を示したフロー図である。

正サイト100のDBMS122は、ユーザからのDB処理(データ更新)要求により、更新するデータをストレージ装置130からDBバッファ404に読み出してデータを更新する(ステップ1102)。

【0064】

続いてDBMS 122は、当データ更新を識別するために付与したLSN、ログの種別、更新されたデータのデータ領域ID、データのアドレス、更新データを一つのログとしてログバッファ402に書き込む（ステップ1104）。

【0065】

続いてDBMS 122は、ログバッファ402の記憶領域に空きが無くなったかどうか判定する（ステップ1106）。ログバッファ402の記憶領域に空きが無い場合、DBMS 122は、ログバッファ402内のログデータを、データ領域500のうち、ログ格納用に作成したデータ領域に書き込む（ステップ1108）。

【0066】

この際、ログ格納用のデータ領域のデータ優先度は0であるので、ストレージ装置130は、RCプログラム140を実行して、ログデータを副サイト102へ同期リモートコピーする。尚、リモートコピー処理1200については後述する（ステップ1110）。

【0067】

続いて、DBMS 122は、DBバッファ404の記憶領域に空きがあるかどうか判定する（ステップ1112）。DBバッファ404に空きが無くなった場合、DBMS 122は、DBバッファ404内の更新されたデータを、対応するデータ領域の所定のアドレスで示される場所に書き込む（ステップ1114）。又、DBMS 122は、データ領域管理領域502のLSN512に、書き込んだ更新データに対応するログのLSNを格納する（ステップ1116）。

【0068】

なお、ログデータ及びDBデータのデータ領域への書き込み契機をログバッファ402及びDBバッファ404の記憶領域に空きが無くなった時で説明したが、各バッファの空きの有無ではなく、ある一定期間毎といった別の契機でDBMS 122がデータ領域への書き込みを行ってもかまわない。

【0069】

図12は、RCプログラム140によるリモートコピー処理1200の手順の例を示したフロー図である。

正サイト100のRCプログラム140は、データ優先度が0のグループに属するLUのデータが更新された場合（ステップ1202）、更新されたデータを副サイト102に送信し（ステップ1204）、副サイト102からの応答を待つ（ステップ1206）。正サイト100のRCプログラム140は、副サイト102からの応答を受信して次の処理に移る。

【0070】

一方、正サイト100でのデータ更新がデータ優先度0以外の場合又はデータ更新が発生しない場合、正サイト100のRCプログラム140は決められたある一定時間の経過を待つ（ステップ1208）。一定時間経過後、正サイト100のRCプログラム140は、データ優先度が0以外のグループに含まれるLUに対して行われたデータ更新を検索し、その更新されたデータをグループのデータ優先度順で副サイト102に送信する（ステップ1210）。

【0071】

図13は、副サイト102の運用管理プログラム120による災害監視処理の手順例を示したフロー図である。

副サイト102の運用管理プログラム120は、正サイト100の運用管理プログラム120に対し、定期的にはアライブ通信を行う。アライブ通信とは、お互いがお互いを監視しあうために遣り取りされる通信で、一般的にはハートビート信号等が使用される（ステップ1302）。

【0072】

ステップ1302のアライブ通信に対する応答が正サイト100から帰ってきた場合、副サイト102は、正サイト100が正常に動作していると判定し、一定期間をおいてステップ1302からの処理を繰り返す。

【0073】

一方、正サイト100から応答がなかった場合、副サイト102は、正サイト100に障害、もしくは災害が発生したと判定する。そして、副サイト102の運用管理プログラム120は、副サイト102のDBMS122を起動して運用を開始するとともに（ステップ1306）、DBMS122は、DB回復処理1400を実行する。なお、このときDBMS122は、DBデータを記憶する各データ領域を閉塞状態にしておき、データの回復が終わったデータ領域から順次オープンしていく。ここで「閉塞状態」とは、データ領域のデータにDBサーバ110のアプリケーションプログラム等がアクセスができない状態のことを指す。又、「オープン」とは、閉塞状態が解除され、データ領域のデータがアプリケーションプログラム等によって使用可能な状態にすることを指す（ステップ1308）。

【0074】

図14は、副サイト102のDBMS122によるDB回復処理1400の手順例を示したフロー図である。

副サイト102のDBMS112は、データ領域設定情報406を参照して、データ優先度が高いデータ領域からデータの回復を行い（ステップ1402）、データの回復が終わればそのデータ領域をオープンして使用可能状態にする（ステップ1404）

回復の方法の手順は次の通りである。まずDBMS122は、データ領域のデータ領域管理領域502内にある領域512のLSNを取得し、そのデータ領域で一番最後に更新されたデータのログを特定する。そして、DBMS122は、同期リモートコピーでコピーされたログデータから、LSNで特定したログより後のログの中でそのデータ領域のデータを更新したことを示すログがあれば、そのログのエントリに格納された更新データを用いてDBデータの更新を行う。

【0075】

例えば、図7でデータ「A」「D」「C」が副サイト102の各データ領域にコピーされている状態からデータ回復を行う場合、データ優先度が最も高いデータ領域700のLSNは「13」であり、LSNが「13」以降のログデータ710の中でデータ領域700のデータ更新を示しているログはないため、DBMS122は、データ領域700をそのままオープンして使用可能状態にする。

【0076】

次にデータ優先度が高いデータ領域704のLSNは「12」であり、LSNが「12」以降のログデータ710の中でLSNが「15」のログがデータ領域704のデータ更新を示している。従って、DBMS122は、LSNが「15」のログが持つ更新データ「F」をデータ領域704に書き込み、領域512のLSNを「15」に更新した後、データ領域704をオープンして使用可能状態にする。

【0077】

最後に最もデータ優先度が低いデータ領域702のLSN512は「9」であり、LSNが「9」以降のログデータ710の中でLSNが「11」と「14」のログがデータ領域702のデータ更新を示している。従って、DBMS122は、LSNが「11」のログが持つ更新データ「B」と、LSN「14」のログが持つ更新データ「E」をデータ領域702に書き込み、LSN512を「14」に更新した後、データ領域702をオープンして使用可能状態にする。

【0078】

ここで、従来技術と本発明とのデータ回復の処理の違いを図7を用いて説明する。

図7において、更新のあったデータ「A」～「F」が正サイトから副サイトにリモートコピーされる際、従来技術ではデータの更新順で副サイトにデータが転送されるが、本発明では各データが書き込まれるデータ領域の優先度順でコピーされる。つまり、データ「A」～「F」の順番ではなく、優先度が最も高いグループ1のデータ「A」「D」、次に優先度が高いグループ3のデータ「C」「F」、最後に優先度が最も低いグループ2のデータ「B」「E」の順番で副サイトにコピーされる。

【0079】

ここで、4つ目のデータをコピーしている時に、正サイトに障害が発生し、副サイトでDBの復旧を図ると仮定する。従来技術ではデータ「A」、「B」及び「C」が副サイトにコピーされた状態から復旧を図ることになり、ログデータ710からデータ「D」、「E」及び「F」を回復する。つまり、どのデータ領域にも回復が必要なデータが存在する可能性があり、DBデータの復旧時間が延びてしまう。

【0080】

一方、本発明ではデータ「A」、「D」及び「C」が副サイトにコピーされた状態から、優先度の高いグループに属するデータ領域順に復旧を図るため、グループ1に属するデータ領域、グループ3に属するデータ領域、グループ2に属するデータ領域の順番でデータの回復を行う。

【0081】

ここで本発明においては、グループ1に属するデータ領域700のデータ「A」及び「D」は既にコピーされているため、副サイト側のDBMSはログデータ710からDBデータを回復する必要はなく、グループ1に属するデータ領域をそのままオープンして使用可能状態にすることができる。又、グループ3に属するデータ領域704のデータ「C」は既にコピーされているため、副サイト側のDBMSはログデータ710からデータ「F」のみを回復した後、グループ3に属するデータ領域をオープンして使用可能状態にする。一方、グループ2に属するデータ領域702のデータ「B」「E」はまだコピーされていないため、副サイト側のDBMSはログデータ710からデータ「B」及び「E」を回復した後、グループ2に属するデータ領域をオープンして使用可能状態にする。以上の手順により、本発明では、優先度の高いデータ領域の復旧時間を短くすることが可能となる。

【0082】

上述したように、本発明では、優先度を高く設定したデータ領域のデータを優先的に福サイトにコピーし、障害等により福サイト上でDBの回復を図る際は高優先度のデータ領域のデータから復旧を図ることで、優先度の高いデータ領域のデータの復旧時間を短くすることができる。

【0083】

つまり、本発明においては、ある時点において正サイトに障害が発生し、副サイト上でDB復旧が行われる際、復旧優先度の高いデータは復旧優先度の低いデータに比べ、副サイト上にコピーされている可能性が高くなるため、この場合ログからのデータ復旧を行う必要がなく、したがって復旧時間が短くなる。

【図面の簡単な説明】

【0084】

【図1】 計算機システムのシステム構成例を示す図である。

【図2】 制御情報140の構成例を示した図である。

【図3】 ローデバイス情報300の例を示した図である。

【図4】 DBMS122の構成例を示した図である。

【図5】 データ領域の構成例を示した図である。

【図6】 正サイト100から副サイト102へのリモートコピーの概略を示した図である。

【図7】 リモートコピーの概略を示した図である。

【図8】 記憶領域の割当て処理の手順例を示したフロー図である。

【図9】 DBの構築処理の手順例を示したフロー図である。

【図10】 グループ作成処理の手順例を示したフロー図である。

【図11】 データ更新処理の手順例を示したフロー図である。

【図12】 リモートコピー処理の手順例を示したフロー図である。

【図13】 副サイト102の運用管理プログラム120による災害監視処理の手順例を示したフロー図である。

【図 14】副サイト 102 の DBMS 122 による DB 回復処理の手順例を示したフロー図である。

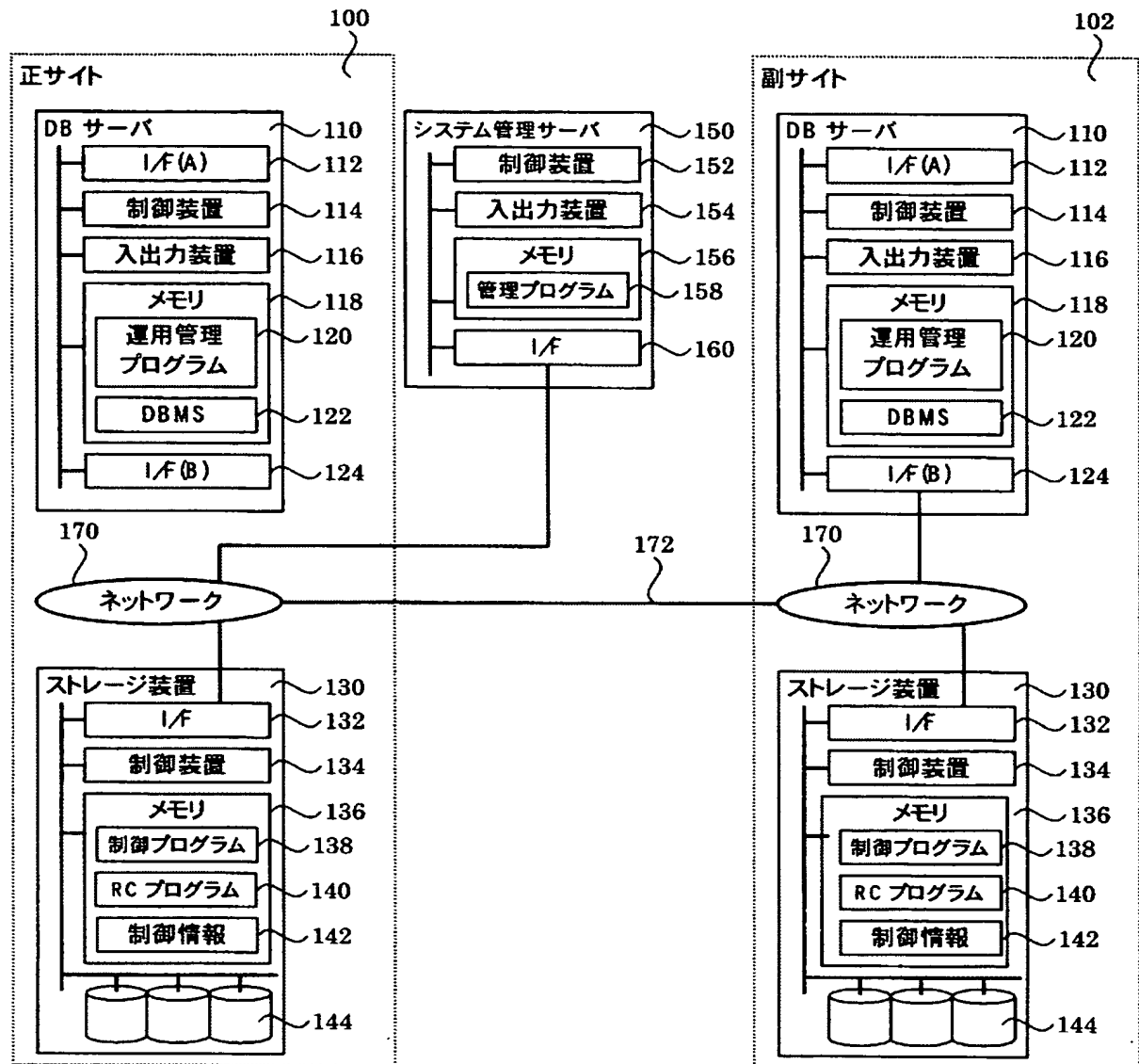
【符号の説明】

【0085】

100…正サイト、102…副サイト、110…DBサーバ、112…I/F (A)、114…制御装置、116…入出力装置、118…メモリ、120…運用管理プログラム、122…DBMS、130…ストレージ装置、134…制御装置、136…メモリ、144…ディスク、150…システム管理サーバ、170…ネットワーク。

【書類名】 図面
【図 1】

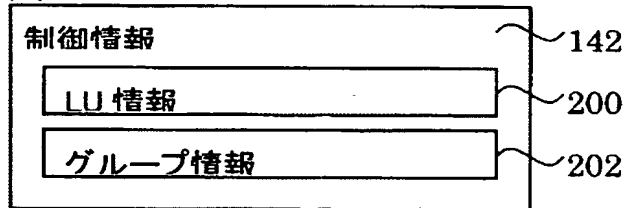
図1



【図 2】

図2

(a)



(b)

LUN	ブロック数	サーバ ID
0	20971520	0
1	419430400	0
2	209715200	0
3	209715200	0
...

Diagram (b) shows a table of LU information. The table is labeled 210 and has three columns: LUN (212), ブロック数 (212), and サーバ ID (214). The table is part of a larger structure labeled 200, which is also labeled 216. The table contains five rows of data, with the last row showing ellipses (...).

(c)

グループ ID	データ 優先度	LUN
0	0	0
1	1	1.4...
2	5	2.5...
3	3	3.6...
...

Diagram (c) shows a table of group information. The table is labeled 220 and has three columns: グループ ID (222), データ優先度 (222), and LUN (224). The table is part of a larger structure labeled 202, which is also labeled 226. The table contains five rows of data, with the last row showing ellipses (...).

【図 3】

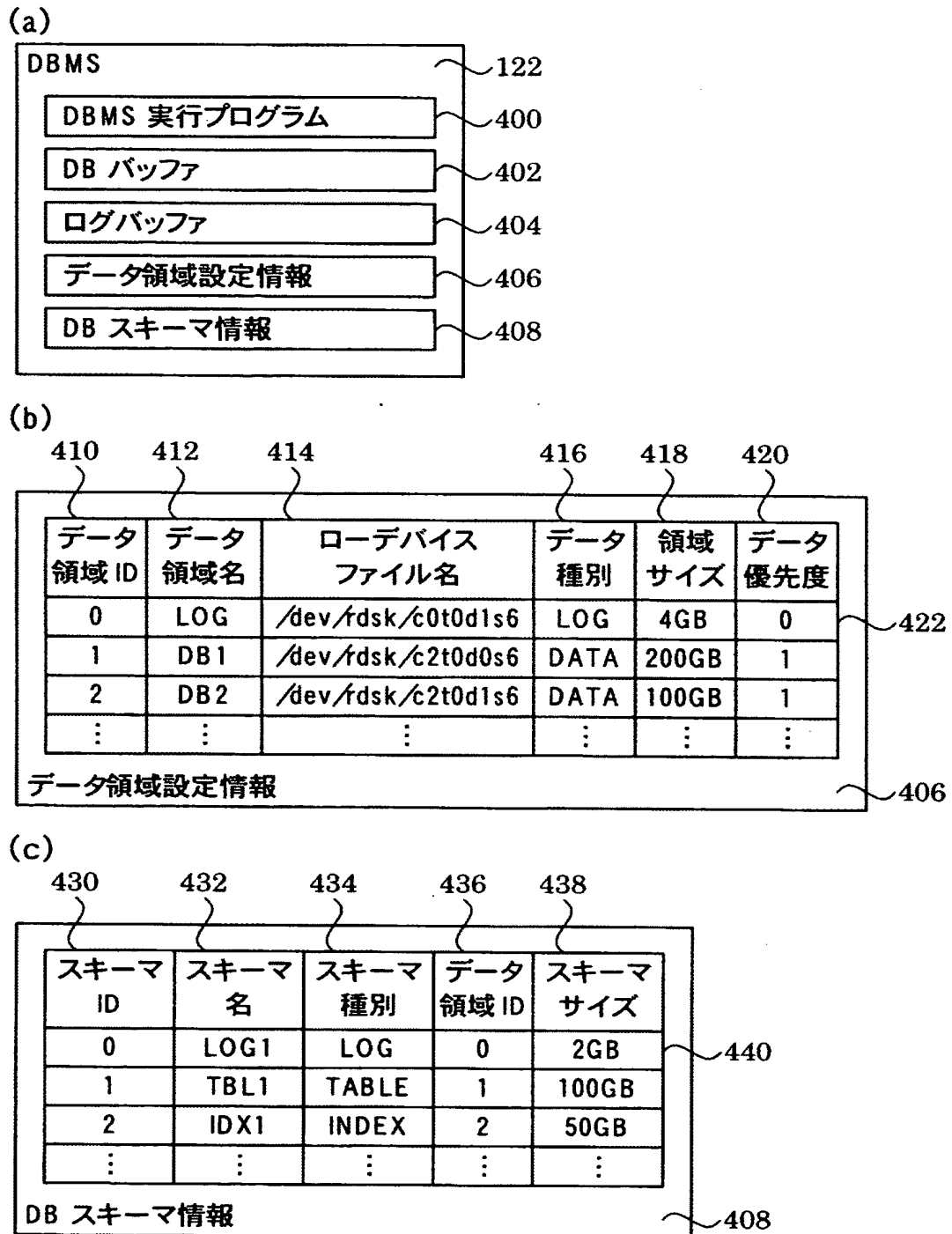
図3

ローデバイス ファイル名	ストレージ 装置 ID	LUN
/dev/fdisk/c0t0d1s6	0	0
/dev/fdisk/c2t0d0s6	0	1
/dev/fdisk/c2t0d1s6	0	2
⋮	⋮	⋮

ローデバイス情報

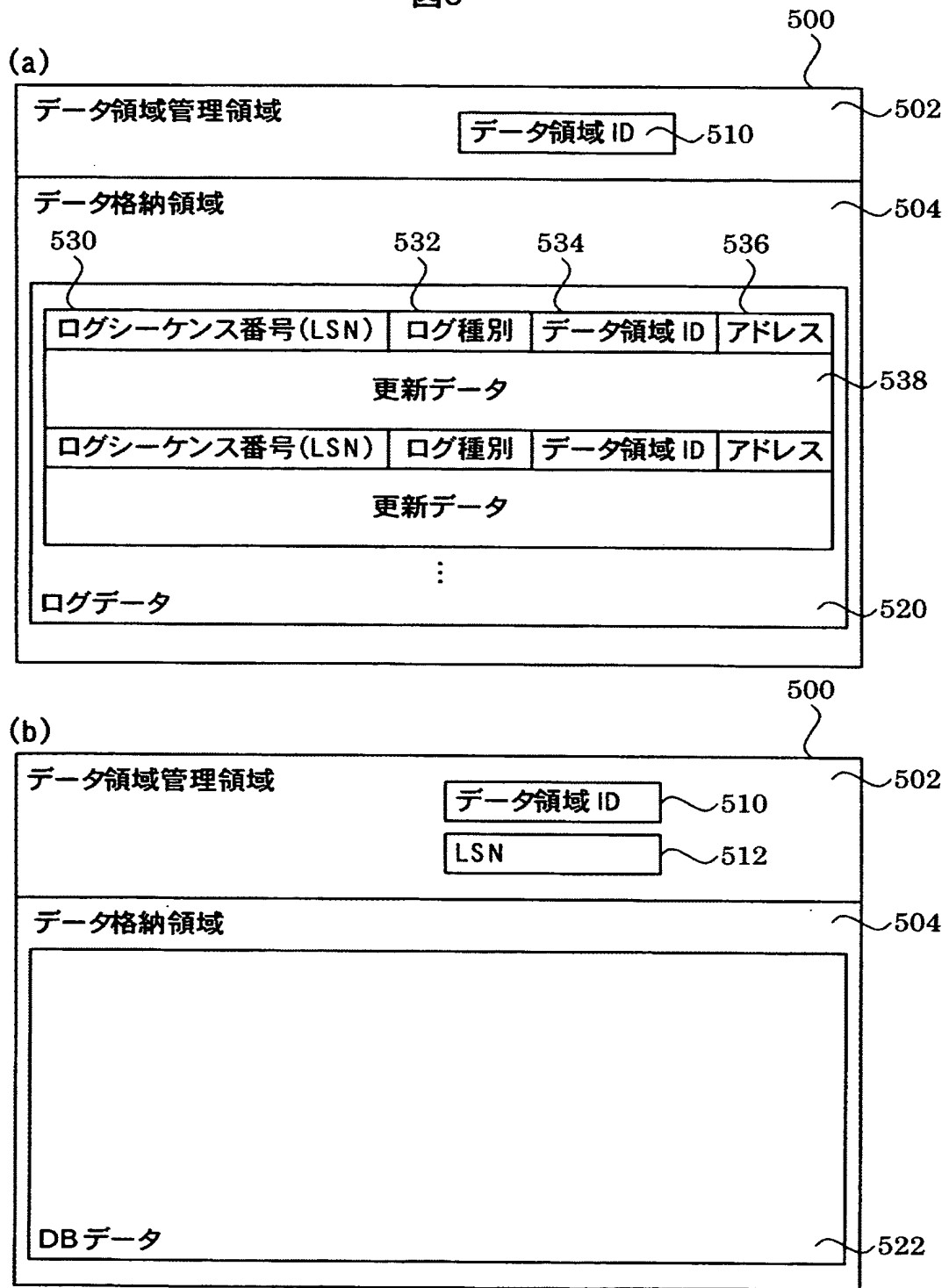
【図 4】

図4



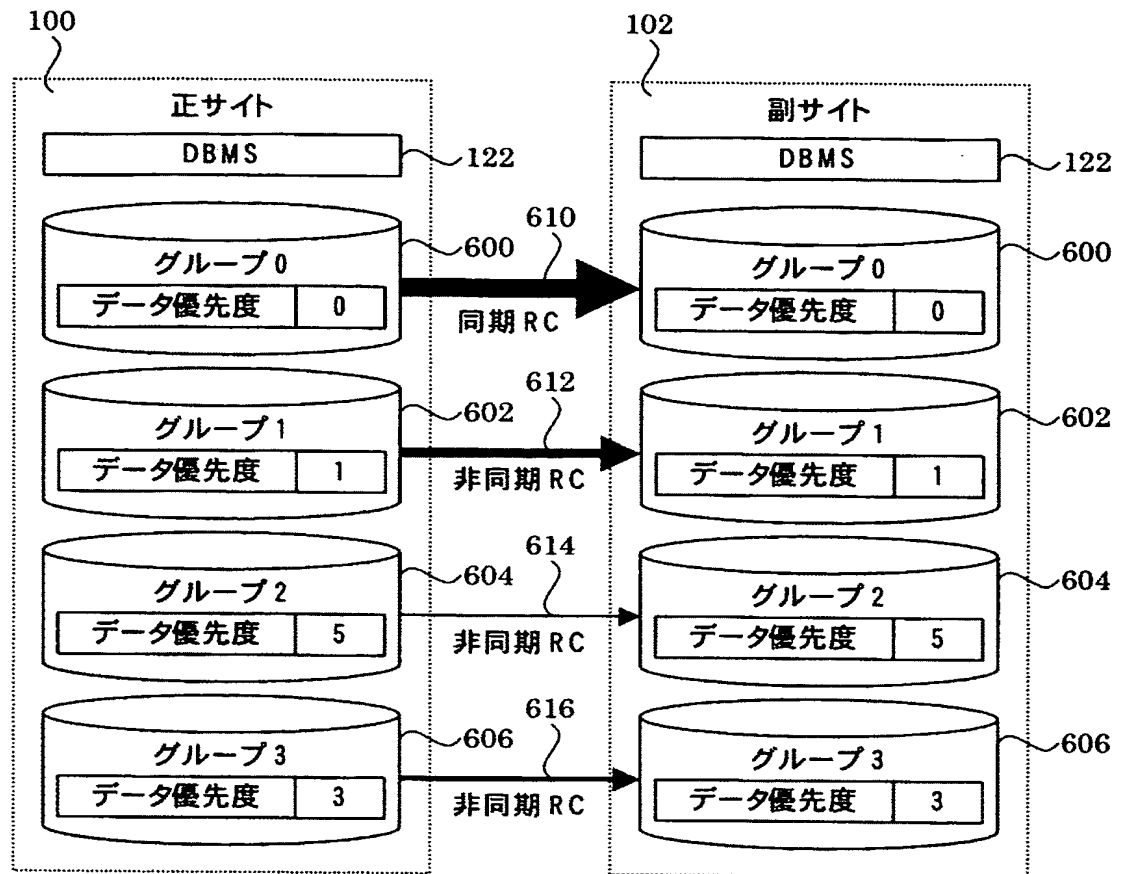
【図 5】

図5

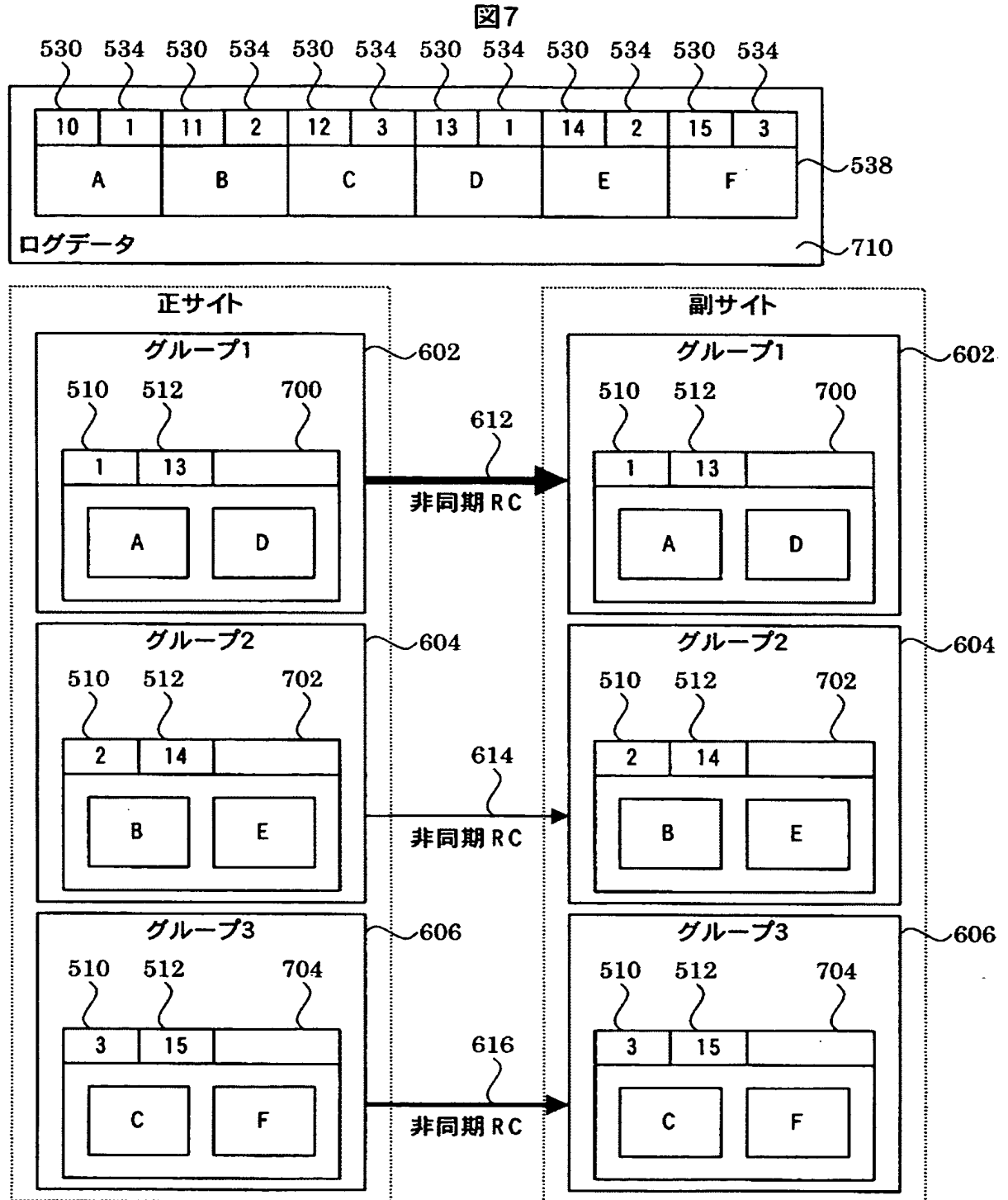


【図 6】

図6

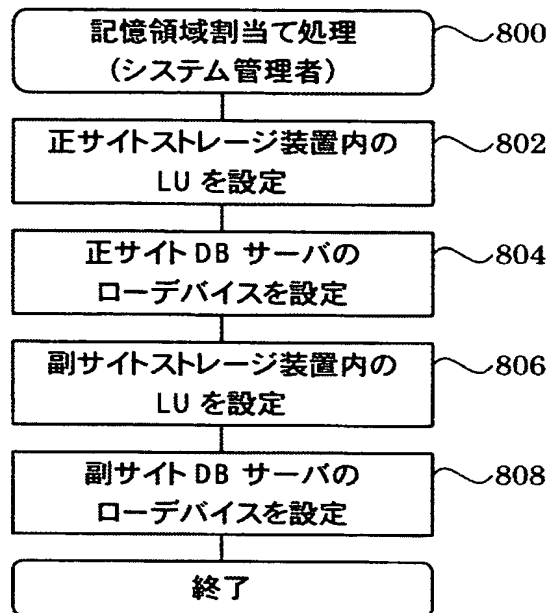


【図 7】



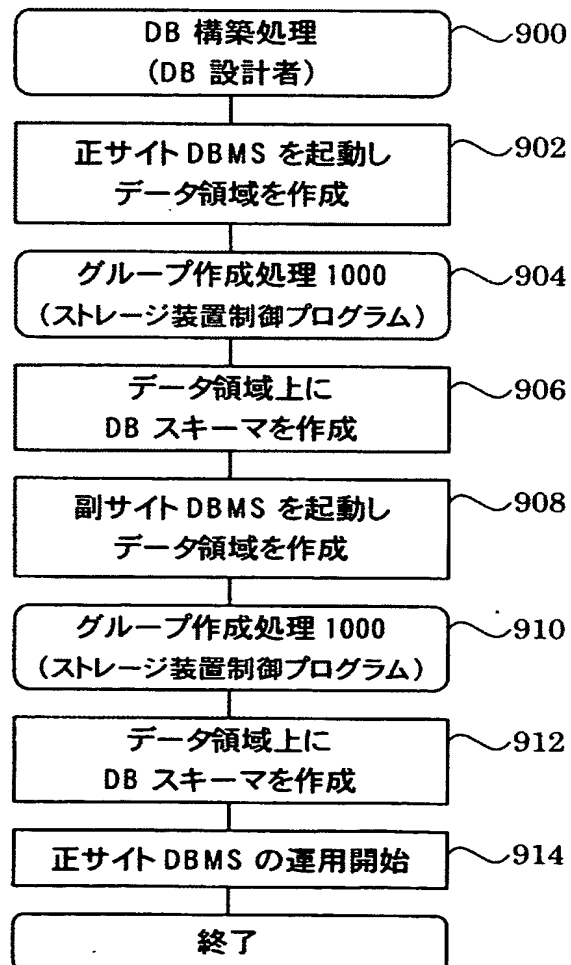
【図 8】

図 8



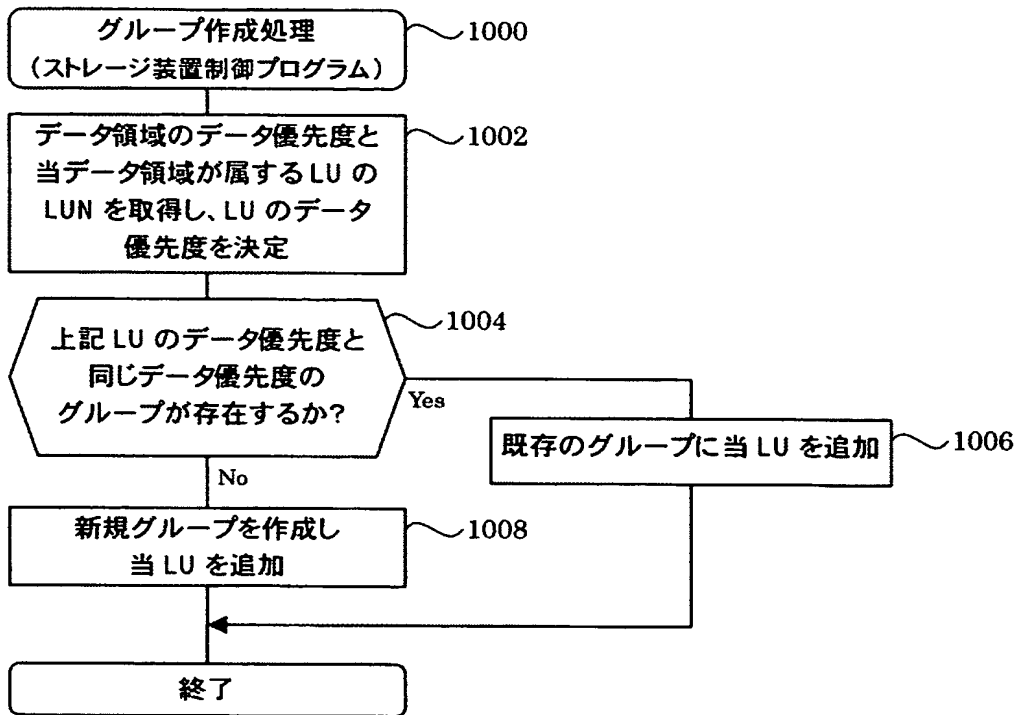
【図 9】

図 9



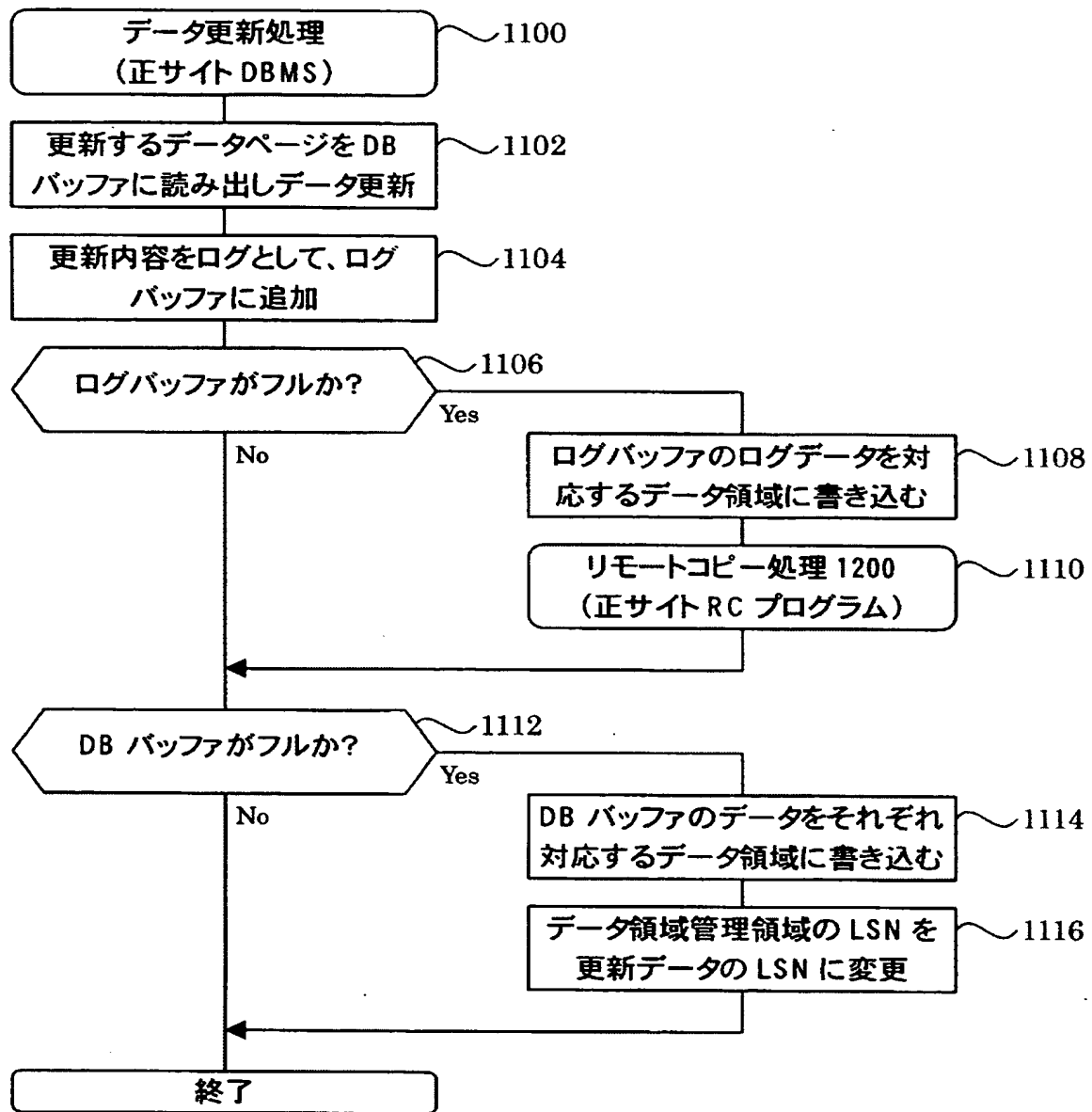
【図 10】

図10

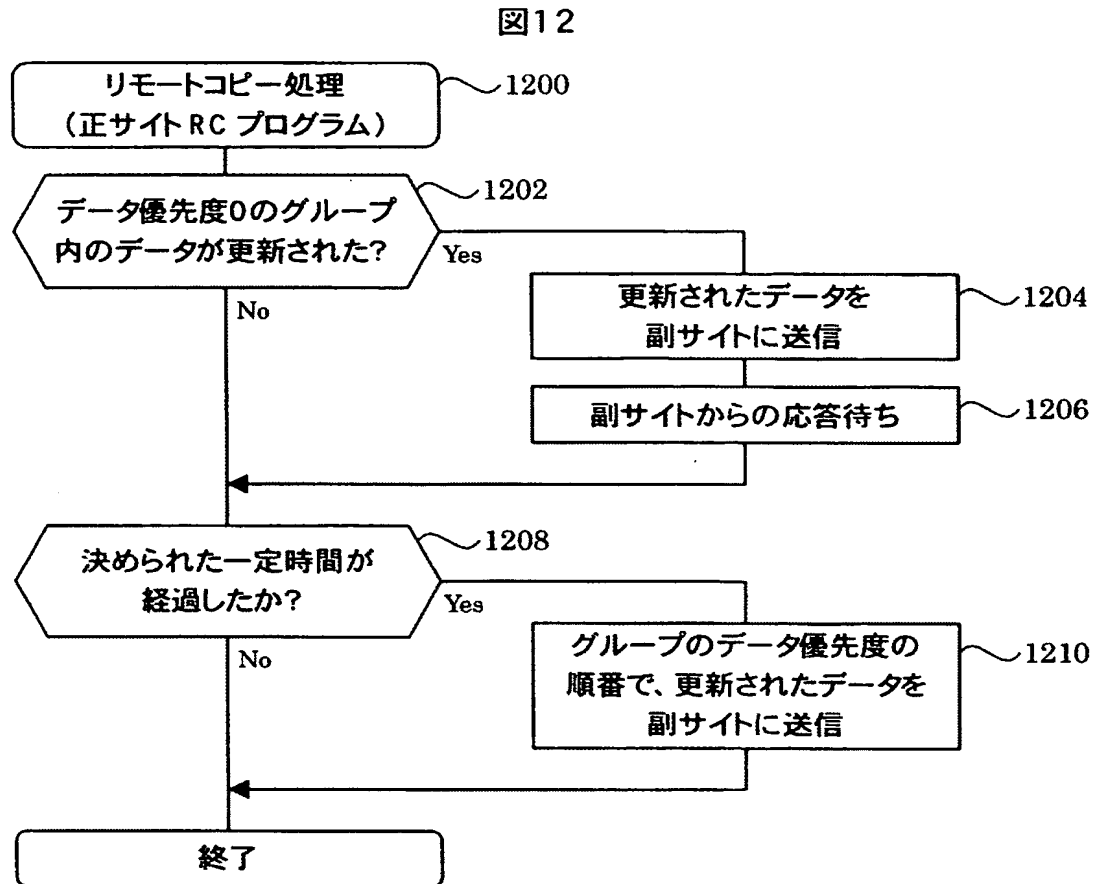


【図 11】

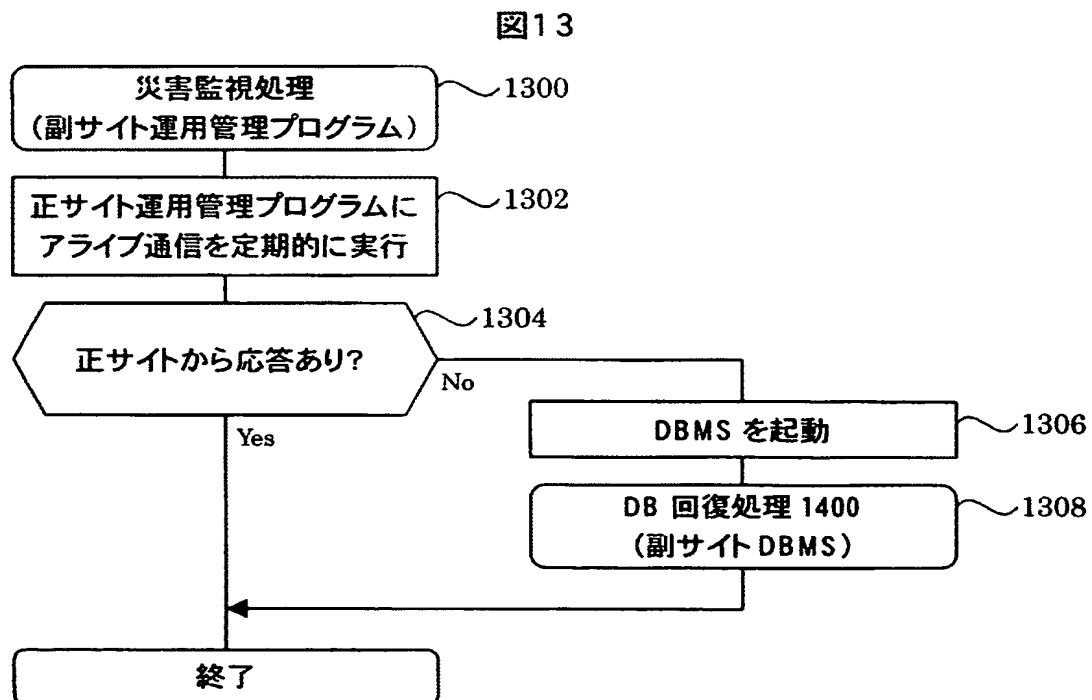
図 11



【圖 12】

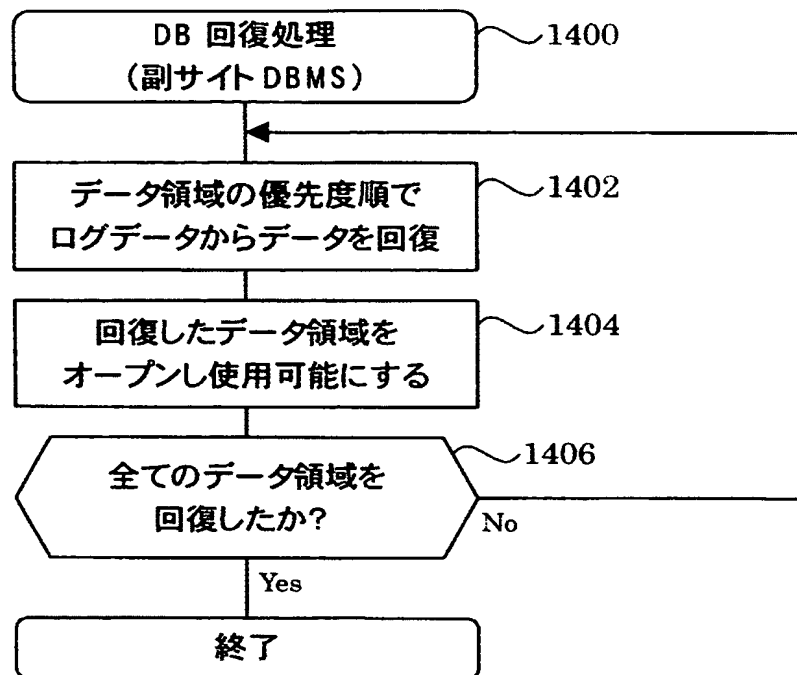


【図 13】



【図 14】

図14



【書類名】 要約書**【要約】****【課題】**

DRシステムとDBを構築する際、DBデータ毎に復旧時間の要求が異なる場合も、従来技術ではそれらを考慮することはなく、最も条件の高いDBデータに対する要求がDRシステム構築の要件となり、システムがオーバースペックになる可能性が高いという課題がある。

【解決手段】

本発明では、ある管理単位毎にDBデータの優先度を設定する手段、そのデータ優先度順に正サイトから副サイトにデータをコピーする手段、正サイトに災害が発生し副サイトでDBの回復を行う際、データ優先度順にデータの回復を行い、順次使用可能な状態にする手段を設ける。

【選択図】 図6

認定・付加情報

特許出願の番号	特願 2004-047176
受付番号	50400290370
書類名	特許願
担当官	末武 実 1912
作成日	平成16年 2月25日

<認定情報・付加情報>

【提出日】 平成16年 2月24日

特願 2 0 0 4 - 0 4 7 1 7 6

出 願 人 履 歴 情 報

識別番号 [0 0 0 0 0 5 1 0 8]

1. 変更年月日	1 9 9 0 年 8 月 3 1 日
[変更理由]	新規登録
住 所	東京都千代田区神田駿河台 4 丁目 6 番地
氏 名	株式会社日立製作所